Medical Image Retrieval Based on Ensemble Learning using Convolutional Neural Networks and Vision Transformers

Y. Ahmed Yahya[1*], D. Khaled[2], W. Khaild Al-Azzawi[3], T. Alghazali[4], H. Sabah Jabr[5], R. Madhat Abdulla[6], M. Kadhim Abbas Al-Maeeni[7], N. Hussin Alwan[8], S. Saad Najeeb[9], Kh. T. Falih[10]

1- Department of nursing, Al-Hadba University College, Iraq.
Email: yahya@hcu.edu.iq (Corresponding author)
2 Al-Manara College for Medical Sciences, Maysan, Iraq.
Email: dalyakhaled@uomanara.edu.iq
3- Medical Lab. Techniques department, College of Medical Technology, Al-Farahidi University, Iraq.
Email: waleed.khalid@alfarahidiuc.edu.iq
4- College of Media, Department of Journalism, The Islamic University in Najaf, Najaf, Iraq.
Email: gazali.tawfeeq@gmail.com
5- Anesthesia Techniques Department, Al-Mustaqbal University College, Babylon, Iraq.
Email: huda.sabah@mustaqbal-college.edu.iq
6- The University of Mashreq, Baghdad, Iraq.
Email: rusul.madhat@uom.edu.iq
7- Al-Nisour University College, Baghdad, Iraq.
Email: mohammed.kadhimaam@gmail.com
8- Department of Nursing, Al-Zahrawi University College, Karbala, Iraq.
Email: natherahussin.alwan@g.alzahu.edu.iq
9- Al-Esraa University College, Baghdad, Iraq
Email: salma@esraa.edu.iq
10- New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar, Iraq.
Email: khaldoon@alayen.edu.iq

Received: April 2022 Revised: June 2022 Accepted: July 2022

ABSTRACT:
The rapid increase in the number of medical image repositories nowadays has led to problems in managing and retrieving medical visual data. This has proved the necessity of Content-Based Image Retrieval (CBIR) with the aim of facilitating the investigation of such medical imagery. One of the most serious challenges that require special attention is the representational quality of the embeddings generated by the retrieval pipelines. These embeddings should include global and local features to obtain more useful information from the input data. To fill this gap, in this paper, we propose a CBIR framework that utilizes the power of deep neural networks to efficiently classify and fetch the most related medical images with respect to a query image. Our proposed model is based on combining Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) and learns to capture both the locality and also the globality of high-level feature maps. Our method is trained to encode the images in the database and outputs a ranking list containing the most similar image to the least similar one to the query. To conduct our experiments, an intermodal dataset containing ten classes with five different modalities is used to train and assess the proposed framework. The results show an average classification accuracy of 95.32 % and a mean average precision of 0.61. Our proposed framework can be very effective in retrieving multimodal medical images with the images of different organs in the body.

KEYWORDS: Content-Based Image Retrieval, Medical Image Retrieval, Ensemble Learning, Convolutional Neural Networks, Vision Transformers, Deep Learning, Similarity-Based Visual Search.

Paper type: Research paper
DOI: https:// 10.30486/mjee.2022.696500
How to cite this paper: Y. A. Yahya, D. Khaled, W. Khaild Al-Azzawi, T. Alghazali, H. S. Jabr, R. M. Abdulla, M. Kadhim Abbas Al-Maeeni, N. H. Alwan, S. S. Najeeb, Kh. T. Falih, **"Medical Image Retrieval Based on Ensemble Learning using Convolutional Neural Networks and Vision Transformers"**, *Majlesi Journal of Electrical Engineering*, Vol. 16, No. 3, pp. 1-7, 2022.

1. INTRODUCTION

With the advent of enormous data centers and storage systems, a large amount of visual content is being created on a daily basis [1]. These advancements have also been significantly beneficial in the context of clinical and diagnostic studies [2]. That is, hospitals and healthcare centers, which normally possess imaging facilities, produce a considerable amount of visual data daily and store them in the databases. These image collections lead to unavoidable conundrums, which hinder the process of managing, searching, and retrieving these visual content [3]. Therefore, developing effective image-based retrieval systems is essential to help clinicians investigate and browse such a huge amount of medical imagery. A reliable solution can be Content-Based Image Retrieval (CBIR) systems which provide an automatic approach for seeking, finding, and fetching relevant images. Without the existence of such systems, the procedure of investigating the target images can be extremely cumbersome and tedious. It is worthwhile to note that textual information such as tags and annotations is not efficient in most cases since it requires expert human resources and time [4]. Thus, CBIR systems can provide a supplementary way to automatically extract representations from the images without the demand for manual work [1].

CBIR is a task related to the domain of Computer Vision, in which we pave the way for searching for visual content relevant to a query image [5]. This search is based on multiple factors such as texture, color, geometry, and other features that can be inferred from an image [6]. In CBIR systems, the input image is first mapped to a high-dimensional feature in feature space [7]. These features are called embeddings, which represent the images in latent space. Then, the embeddings are fed to a similarity-defining module which measures the relevance between the query image and the input one [8]. The performance of these systems is thoroughly dependent on the representational quality of the embeddings. Hence, the more representative these embeddings' principal features, the more relevant images can be retrieved [9].

Hitherto, a considerable number of research works based on machine learning and deep learning-based approaches have been proposed in the literature. Rahman et al. [10] designed a system for CBIR, in which probabilistic multi-class SVM and fuzzy C-mean (FCM) clustering were employed with the aim of reducing the search space. Also, some local feature descriptors have been used for the retrieval task. In [11], Scale Invariant Feature Transform (SIFT) is used to recognize objects and fetch them. Speeded up Robust Features (SURF), in [12], is used to improve the invariance biase and the efficiency of the features to be used for retrieval.

Moreover, in [13], Kundu et al. put forward a CBIR system using pulse coupled neural network and non-rate. Pogarell et al., in [14], used a CNN-based classifier for classifying and retrieving interstitial lung diseases. Their customization of their proposed model's architecture enabled their approach to efficiently extract low-level biased features.

Although the previous works have presented considerable efficacy, the fact that these approaches fail to capture both local and global features, ranging from higher to lower feature maps, is assumed to be a disadvantage. The main problem is that in a majority of cases, both of these types of features are required to extract finer meaningful representations from the patterns existent in the input data [15].

In the current study, we propose a framework that learns both local and global feature representation for learning the embeddings to overcome the previous works' shortages. In the suggested framework, a representation learning model, which comprises CNN and ViT-based models, is assembled together in order to capture better embeddings in terms of representational quality. The principal contributions of this study are as follows:

1. We propose a framework for retrieving medical images that learns both the input data's global and local features.
2. To our knowledge, this study is the first to use ViT-based and CNN-based models for representation learning in a medical image analysis context.
3. This study’s case is in a challenging real-world data gathered from hospitals, and its results are more trustworthy in real scenarios

2. MATERIAL AND METHODS

2.1. Overview

In this section, we present an overview of our proposed framework. As is depicted in Fig. 1, our method comprises two parts. The first one is a representation learning module containing two ensemble models, detailed in section 2.6, and the second one is the similarity score extractor module which is utilized to calculate the similarity of an image with another one. This module recognizes a ranking list of images that are most similar to the query image and fetches them from the database.

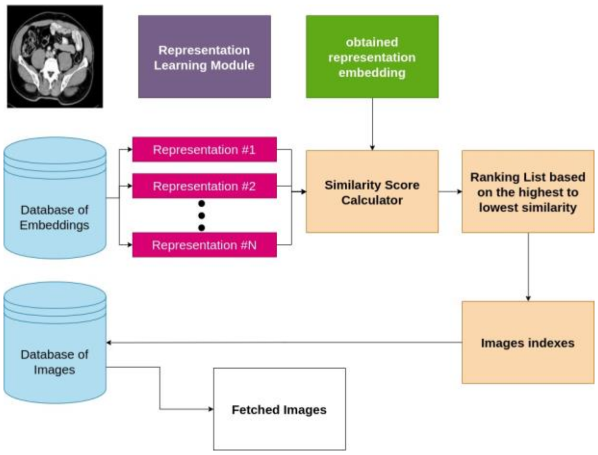
Representation Learning Module
obtained representation embedding
Database of Embeddings
Representation #1
Representation #2
Representation #N
Similarity Score Calculator
Ranking List based on the highest to lowest similarity
Images indexes
Database of Images
Fetched Images

Fig. 1. The overview of the proposed framework.

2.2. Preprocessing

Many researchers consider pre-preprocessing as one of the most elemental stages in any machine learning-based algorithm [16]. It has an exceptional influence on the prediction that a model outputs as its final decision [17]. In this paper, as we deal with decision-making in a deep learning-based model, we have followed the same rule and normalized the input images' pixels to have a specific mean and standard deviation. The chosen values for these two parameters are detailed in Table 7.

2.3. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are ubiquitous in the field of computer vision due to their ability to handle visual data very performantly [18]. These networks have been used in various tasks such as image classification [19], image segmentation [20], object detection [21], activity recognition [22], face recognition [23], video processing [24], etc. Their advantage is the ability to harness the spatial and temporal correlation in the structure of the data [25]. The CNNs' topology is usually sectioned into several learning stages, including convolutional layers, non-linear processing units, named activation functions, and subsampling layers [26]. Fig. 2 shows a typical architecture of a CNN.

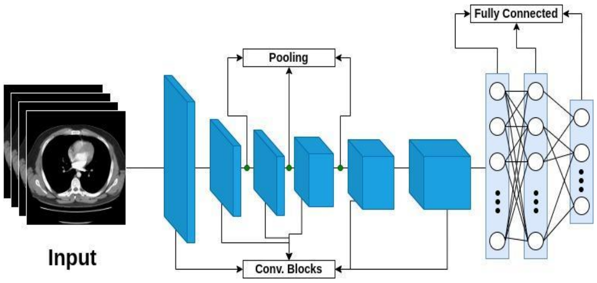

Fig. 2. A common CNN - Figure is the redesign of AlexNet Architecture.

As seen in Fig. 2, after the input data is fed to the network, several blocks of convolutional layers are applied to the image. Each one generates levels of feature maps, learning higher to lower-level features from the data. Finally, the final feature map is fed to a classification head with several fully connected layers. CNN internals contain kernels/filters of settled measurements, and these are alluded to as including finders. Once highlights from a picture are recognized, the data with respect to the position of entities in the input picture can really be neglected [26]. Sub-sampling could be a strategy formulated to decrease the dependence on exact situating inside highlight maps created by convolutional layers inside a CNN [27].

2.4. Vision Transformers

The origin of Vision Transformers (ViT) [26] is the transformer-based models which were introduced in the field of Natural Language Processing (NLP). These models are significantly strong in learning long dependencies of the sequences within the input data [28]. Their success in a variety of tasks, such as speech recognition [29], machine translation [30], text summarization [31], and intelligent chatbots [32], has motivated the researchers to adjust this technique in computer vision. At the core of these models is the self-attention [33] mechanism. This module allows the inputs to interact and learn what should be attended to. The output of the self-attention module is the amalgamation of these interactions and attention scores [34]. Fig. 3 demonstrates a common ViT-based model and its transformer block.

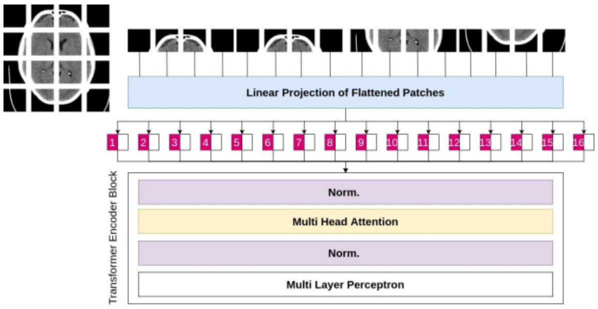

Fig. 3. The architecture of a ViT.

As is witnessed in Fig. 3, the image input is first patchified into multiple parts, with the object being similar to sequential data. Then, these patches are projected into a linear layer combined with positional encodings, which store the position-related information about each part of the image. Next, these embeddings are the input of the transformer block. This block comprises multi-head attention modules and employs the self-attention mechanism and normalization layers. Finally, the output is given to a Multi-Layer Perceptron (MLP) head which gives us the probability distribution in which the predicted class owns the highest probability.

Moreover, in comparison with the CNN-based models, ViT-based architectures have some advantages. Firstly, CNNs heavily depend on their domain and cannot be made domain agnostic [33]. Secondly, CNN lacks the capacity required to learn any global understanding of the image and does not recognize the structural dependency between its features [34].

2.5. Proposed Representation Learning Module

This section includes our proposed method for medical image retrieval. Fig 4. demonstrates an overview of the approach.

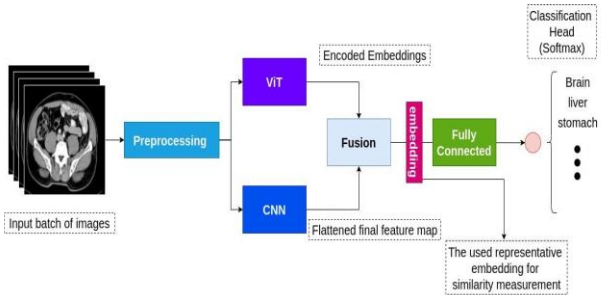
Classification Head (Softmax)
Encoded Embeddings
ViT
Brain
liver
stomach
Preprocessing
Fusion
embedding
Fully Connected
CNN
Input batch of images
Flattened final feature map
The used representative embedding for similarity measurement

Fig. 4. The proposed representation learning module.

As is seen in Fig. 4, the input image is firstly preprocessed by a normalization process on the pixel values. Then, the normalized images are fed to the ViT-based and CNN-based models. The encoded embeddings of the former and the last layer feature maps generated by the latter are fused together and fed to a fully connected block to be classified using a softmax classifier. This way, the network learns to generate the most suitable representations, which can be classified into the correct class in case of being fused. The quality of the generated embedding, which in the figure is displayed by the pink triangle, is of great importance because they will be later, in the inference phase, used to calculate the similarity scores between the different images in the database.

3. RESULTS AND DISCUSSION

3.1. Experimental Setup

The tools detailed in Table 1 are used for implementing the proposed methodology. We trained the proposed ensemble model for 200 epochs for the training phase. Further, the hyperparameters are shown in Table 2. For the training phase, we trained the proposed ensemble model for 200 epochs. Further, the hyperparameters are shown in Table 2.

Table 1. The experimental setup.

Programming language	Python 3.7
Deep learning framework	Pytorch 1.19
CPU	Intel® Core™ i7-10700 CPU @ 2.90GHz × 16
GPU	GeForce GTX 1070

Table 2. Hyperparameters.

Image size	256
Batch size	128
Normalization Standard Deviation	0.229
Normalization Mean	0.485
Optimizer	Adam
Initial learning rate for optimizer	0.0001
Patch size (width, height)	(8, 8)
Number of attention heads	8
Head dimension	128
Drop-out rate for transformer module	0.3
Loss function	Binary Cross Entropy
Last layer	Sigmoid

3.2.Dataset

This section includes a description of the dataset used to assess the proposed methodology. We have gathered a collection of 10000 medical images provided by two well-known hospitals, namely Saint Raphael (Al Rahibat) Hospital and Al Khayal Private Hospital, in Baghdad, Iraq. These images are from 10 different human body organs, for each of which there exist 1000 images. Fig. 5 illustrates one sample for each class. All images are in DICOM (Digital Imaging and COmmunication in Medicine) format, and in our approach, we resize all of them to be in 256 widths and

256 heights. We divide each class in a ratio of 80 to 20 for training and testing sets.

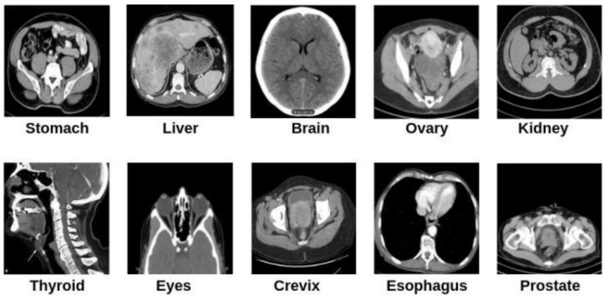

Fig. 5. Samples from each class in the dataset.

3.3. Retrieval performance

Precision and recall are used to assess the proposed methodology's performance. Equation 1 and equation 2 show how these are calculated, respectively.

Table 3 shows our proposed method achieved Average Precision (AP) and Average Recall (AR). Also, the distribution of the data is demonstrated in this table. The training and validation accuracy and loss curves are depicted in Fig. 6.

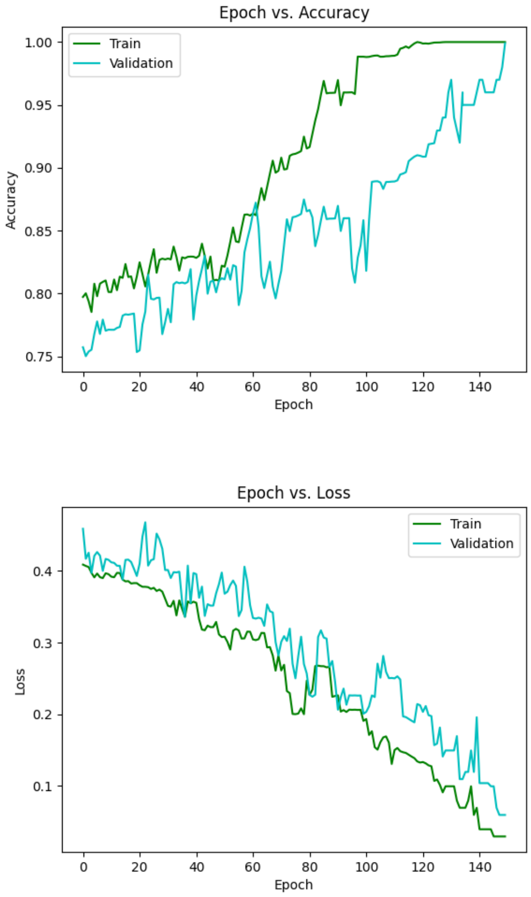

Fig. 6. The epoch vs. accuracy and epoch vs. loss curves.

Moreover, Fig. 7 illustrates two fetched images for three images belonging to three different classes within the dataset. Also, Table 4 demonstrates a comparison between our proposed methodology and the other methods.

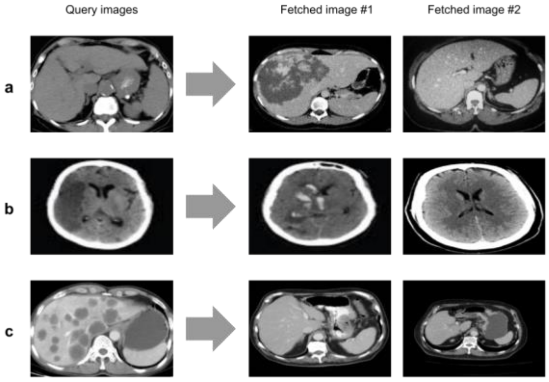

Fig. 7. Some fetched samples for three classes of a) liver, b) brain, c) stomach.

Based on Table 4, it can be inferred that our approach achieves satisfactory results. The achieved value mAP for our proposed methodology is 0.61, which is the second-highest score amongst the other state-of-the-art works. Fig. 8 shows this comparison as well.

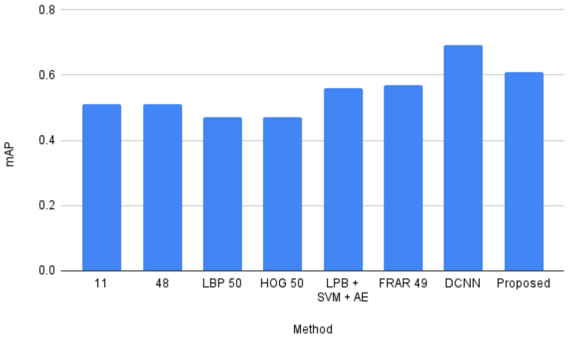

Fig. 8. Comparison between the proposed method and other state-of-the-art works.

4. CONCLUSION

In this paper, an approach for medical image retrieval is proposed. Our method uses two models based on CNN and ViT architectures for the representation learning stage. These representations, which are called embeddings, are used for similarity score indexing between the images in the database and query images. The proposed framework is extensively evaluated on a dataset with 10000 images belonging to 20 various classes of human body organs.

REFERENCES

[1] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK, "**Medical image analysis using convolutional neural networks: a review, "** *Journal of medical systems*. Vol. 42(11), pp.1-3, 2018.

[2] Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, Foran D, Do N, Golemati S, Kurc T, Huang K, **"AI in medical imaging informatics: current challenges and future directions,"** *IEEE Journal of Biomedical and Health Informatics*. Vol. 24(7), pp. 1837-57, 2020.

[3] Li X, Liu S, Lu R, Khan MK, Gu K, Zhang X, **"An efficient privacy-preserving public auditing protocol for cloud-based medical storage system,**" *IEEE Journal of Biomedical and Health Informatics*. Vol. 26(5), pp. 2020-31, 2020.

[4] Latif A, Rasheed A, Sajid U, Ahmed J, Ali N, Ratyal NI, Zafar B, Dar SH, Sajid M, Khalil T, **"Content-based image retrieval and feature extraction: a comprehensive review,"** *Mathematical Problems in Engineering*. 2019.

[5] Li X, Yang J, Ma J, **"Recent developments of content-based image retrieval (CBIR),"** *Neurocomputing*. Vol. 452, pp. 675-89, 2021.

[6] Carvalho ED, Antonio Filho OC, Silva RR, Araujo FH, Diniz JO, Silva AC, Paiva AC, Gattass M, **"Breast cancer diagnosis from histopathological images using textural features and CBIR,"** *Artificial intelligence in medicine*. Vol.105,101845, 2020.

[7] Bressan RS, Bugatti PH, Saito PT, **"Breast cancer diagnosis through active learning in content-based image retrieval,"** *Neurocomputing*. 10, Vol. 357, pp. 1-0, 2019.

[8] Haji MS, Alkawaz MH, Rehman A, Saba T, **"Content-based image retrieval: A deep look at features prospectus,"** *International Journal of Computational Vision and Robotics*. Vol. 9(1), pp. 14-38, 2019.

[9] Tzelepi M, Tefas A, **"Deep convolutional learning for content-based image retrieval,"** *Neurocomputing*. Vol. 275, pp. 2467-78.

[10] Rahman MM, Antani SK, Thoma GR, **"A learning-based similarity fusion and filtering approach for biomedical image retrieval using SVM classification and relevance feedback,"** *IEEE Transactions on Information Technology in Biomedicine*. Vol. 15(4), pp. 640-6, 2011.

[11] Lowe DG, **"Object recognition from local scale-invariant features,**" *InProceedings of the seventh IEEE international conference on computer vision*, Vol. 2, pp. 1150-1157, Ieee, 1999.

[12] Bay H, Ess A, Tuytelaars T, Van Gool L, **"Speeded-up robust features (SURF),"** *Computer vision and image understanding*. Vol. 110(3), pp. 346-59, 2008.

[13] Yonekawa M, Kurokawa H, **"The content-based image retrieval using the pulse-coupled neural network,"** *In The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, IEEE, 2012.

[14] Pogarell T, Bayerl N, Wetzl M, Roth JP, Speier C, Cavallaro A, Uder M, Dankerl P, **"Evaluation of a Novel Content-Based Image Retrieval System for the Differentiation of Interstitial Lung Diseases in CT Examinations,"** *Diagnostics*. Vol. 11(11), pp. 2114, 2021.

[15] Karnila S, Irianto S, Kurniawan R, **"Face recognition using content-based image retrieval for intelligent security,"** *International Journal of Advanced Engineering Research and Science*. Vol. 6(1), pp. 91-8, 2019.

[16] Ha TN, Lubo-Robles D, Marfurt KJ, Wallet BC, **"An in-depth analysis of logarithmic data transformation and per-class normalization in machine learning: Application to unsupervised classification of a turbidite system in the Canterbury Basin, New Zealand, and supervised classification of salt in the Eugene Island mini basin, Gulf of Mexico,"** *Interpretation*. Vol. 9(3), T685-710, 2021.

[17] Raju VG, Lakshmi KP, Jain VM, Kalidindi A, Padma V, **"Study the influence of normalization/transformation process on the accuracy of supervised classification,"** *In2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 729-735, IEEE, 2020.

[18] Li Z, Liu F, Yang W, Peng S, Zhou J, **"A survey of convolutional neural networks: analysis, applications, and prospects,**" *IEEE transactions on neural networks and learning systems*. 2021.

[19] Sarvamangala DR, Kulkarni RV, **"Convolutional neural networks in medical image understanding: a survey,**" *Evolutionary intelligence*. pp.1-22, 2021.

[20] Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D, **"Image segmentation using deep learning: A survey,"** *IEEE transactions on pattern analysis and machine intelligence*. 2021.

[21] Zou Z, Shi Z, Guo Y, Ye J, **"Object detection in 20 years: A survey,"** *arXiv preprint arXiv*:1905.05055. 2019.

[22] Dang LM, Min K, Wang H, Piran MJ, Lee CH, Moon H, **"Sensor-based and vision-based human activity recognition: A comprehensive survey,"** *Pattern Recognition*. Vol. 108, 107561, 2020.

[23] Adjabi I, Ouahabi A, Benzaoui A, Taleb-Ahmed A, **"Past, present, and future of face recognition: A review,"** *Electronics*.Vol. 9(8), pp.1188, 2020.

[24] Hussain T, Muhammad K, Ding W, Lloret J, Baik SW, de Albuquerque VH, **"A comprehensive survey of multi-view video summarization,"** *Pattern Recognition*. Vol. 109, pp. 107567, 2021.

[25] Tan M, Le Q, **"Efficientnet: Rethinking model scaling for convolutional neural networks,"** In *International conference on machine learning*, pp. 6105-6114, PMLR, 2019.

[26] Khan A, Sohail A, Zahoora U, Qureshi AS, **"A survey of the recent architectures of deep convolutional neural networks,"** *Artificial intelligence review*. Vol. 53(8), pp. 5455-516, 2020.

[27] Nejatian S, Parvin H, Faraji E, **“Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification,”** *Neurocomputing*. Vol. 276, pp. 55-66, 2018.

[28] Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, Feichtenhofer C, **“Multiscale vision transformers,”** *InProceedings of the IEEE/CVF International Conference on Computer Vision 2021* pp. 6824-6835, 2021.

[29] Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K, **“Speech recognition using deep neural networks: A systematic review,”** *IEEE Access*. Vol. 7, pp. 19143-65, 2019.

[30] Dabre R, Chu C, Kunchukuttan A, **“A survey of multilingual neural machine translation,**” *ACM Computing Surveys (CSUR)*. Vol. 53(5), pp. 1-38, 2020.

[31] El-Kassas WS, Salama CR, Rafea AA, Mohamed HK, **“Automatic text summarization: A comprehensive survey,”** *Expert Systems with Applications*. Vol. 165, pp. 11367, 2021

[32] Almansor EH, Hussain FK, **“Survey on intelligent chatbots: State-of-the-art and future research directions,**” *In Conference on Complex, Intelligent, and Software Intensive Systems*, pp. 534-543, Springer, Cham, 2019.

[33] Zhao H, Jia J, Koltun V, **“Exploring self-attention for image recognition,**” *InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076-10085, 2020.

[34] Chen J, Ho CM, **“MM-ViT: Multi-modal video transformer for compressed video action recognition,**” *InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2022*, pp. 1910-1921, 2022.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, **"Imagenet classification with deep convolutional neural networks,"** *in Advances in neural information processing systems*, pp. 1097-1105, 2012.

[36] K. He, X. Zhang, S. Ren, and J. Sun, **"Deep residual learning for image recognition,"** *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[37] M. Subrahmanyam, R. Maheshwari, and R. Balasubramanian, **“Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking,”** *Signal Processing*, Vol. 92, No. 6, pp. 1467-1479, 2012.

[38] K. Velmurugan and L. D. S. S. Baboo, **“Image retrieval using harris corners and histogram of oriented gradients,”** *International Journal of Computer Applications (0975-8887)*, Vol. 24, 2011.

[39] K. He, X. Zhang, S. Ren, and J. Sun, **"Deep residual learning for image recognition,"** *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, , pp. 770-778, 2016.

[40] P. Sermanet et al., **“Overfeat: Integrated recognition, localization, and detection using convolutional networks,”** *arXiv preprint arXiv*:1312.6229, 2013.